BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

February 17, 2000

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: *60/107,293*
FILING DATE: *November 09, 1998*

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

T. WALLACE
Certifying Officer

A Bov

PROVISIONAL PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Docket No:29314/34158

PROVISIONAL PATENT APPLICATION TRANSMITTAL

Box Provisional Patent Application
Assistant Commissioner for Patents
Washington, D.C. 20231

Sir:

Transmitted herewith for filing is the provisional patent application under 37 CFR 1.53(c) of

Inventors: Marc Zabeau
Patrick Stanssens

Residences: Onafhankelijkheidslaan 38, 9000 Gent, Belgium
Constant Permekelaan 48, 9830 St.-Martens Latem, Belgium

Title: RESTRICTED AMPLICON ANALYSIS

1. Application Papers Enclosed

1	Title Page
34	Pages of Specification (excluding Claims, Abstract & Drawings)
0	Pages of Claims
0	Pages of Abstract
4	Sheets of Drawings (Figs. 1 to 4)
☐	Formal
☒	Informal

CERTIFICATION UNDER 37 CFR 1.10

I hereby certify that this Provisional Patent Application Transmittal and the documents referred to as enclosed therewith are being deposited with the United States Postal Service on **November 9, 1998**, in an envelope addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231 utilizing the "Express Mail Post Office to Addressee" service of the United States Postal Service under Mailing Label No. EM 362 728 640 US.

David W. Clough

2. **Additional Papers Enclosed**

- ☐ Declaration of Biological Deposit
- ☐ Computer readable copy of sequence listing containing nucl otide and/or amino acid sequenc
- ☐ Verified statement(s) claiming small entity status under 37 CFR 1.9 and 1.27
- ☐ Associate power of attorney
- ☐ Verified translation of a non-English patent application
- ☐ An assignment of the invention
- ☐ Other (e.g. declaration or oath)

3. **Filing Fee Calculation (37 CFR 1.16)**

☒ **Provisional Application (\$75.00/\$150.00)**	Filing Fee:	\$150.00
☐ **Other Fees (e.g. Recording Assignment)**	Filing Fee:	\$______
	Total Fees Enclosed	\$150.00

4. **Method of Payment of Fees**

☒ Enclosed is our firm check in the amount of:	\$150.00
☐ Charge Deposit Account No. 13-2855 in the amount of: A copy of this Transmittal is enclosed.	\$______
☐ Not enclosed	

5. **Deposit Account and Refund Authorization**

The Commissioner is hereby authorized to charge any deficiency in the amount enclosed or any additional fees which may be required during the pendency of this application under 37 CFR 1.16 or 37 CFR 1.17 or under other applicable rules (except payment of issue fees), to Deposit Account No. 13-2855. A copy of this Transmittal is enclosed.

Please refund any overpayment to Marshall, O'Toole, Gerstein, Murray & Borun at the address below.

Please direct all future communications to David W. Clough, at the address below.

Respectfully submitted,

MARSHALL, O'TOOLE, GERSTEIN,
MURRAY & BORUN
6300 Sears Tower
233 South Wacker Drive
Chicago, Illinois 60606-6402
(312) 474-6300

By: ____________________
David W. Clough
Reg. No: 36,107

November 9, 1998

JOINT INVENTORS

Att rney Docket No. 29314/34158

"EXPRESS MAIL" mailing label No. EM 362 728 640 US.
Date of Deposit: November 9, 1998
I hereby certify that this paper (or fee) is being deposited with the United States Postal Service "EXPRESS MAIL POST OFFICE TO ADDRESSEE" service under 37 CFR §1.10 on the date indicated above and is addressed to: Assistant Commissioner for Patents, Washington, D.C. 20231

David W. Clough

PROVISIONAL APPLICATION FOR UNITED STATES LETTERS PATENT

SPECIFICATION

TO ALL WHOM IT MAY CONCERN:

Be it known that we, Marc Zabeau, a citizen of Belgium, residing at Onafhankelijkheidslaan 38, 9000 Gent, in the Country of Belgium and Patrick Stanssens, a citizen of Belgium, residing at Constant Permekelaan 48, 9830 St.-Martens Latem, in the Country of Belgium, have invented a new and useful RESTRICTED AMPLICON ANALYSIS, of which the following is a specification.

Attorneys for Applicants:

David W. Clough (Reg. No. 36,107)
Marshall, O'Toole, Gerstein,
Murray & Borun
6300 Sears Tower
233 South Wacker Drive
Chicago, Illinois 60606-6402

RESTRICTED AMPLICON ANALYSIS

Background of the Invention

Molecular approaches for genetic analyses trace the nucleotide sequence variation that occurs naturally and randomly in the genomes of all living species. Knowledge of the DNA polymorphisms among individuals and between populations is important in understanding the complex links between genotypic and phenotypic variation. In the absence of complete data about sequence variation, one relies on the ability to identify 'nearby' markets that allow to infer the location of certain relevant loci or causal sequence variations. The informativeness of the marker depends on the magnitude of the linkage disequilibrium. Markers can be used in linkage studies to search for candidate genes and in association studies to identify the functional allelic variation on candidate genes that influence inter-individual variation.

The vast majority of sequence variation consists of nucleotide substitutions, often referred to as single nucleotide polymorphism's (SNPs), resulting from mutations that have accumulated during evolution. Most of these nucleotide changes are genetically silent; i.e., they have no measurable biological effect, but provide an immense reservoir of variation in DNA structure. Most methods for genetic analysis used today rely on the detection of nucleotide sequence variation which can be measured by DNA fragment analysis using electrophoretic separation, in which DNA fragments are fractionated based on size or conformation. Occasionally the nucleotide sequence variation will affect either the presence of the DNA fragment or its mobility. In this way the primary nucleotide sequence variation will give rise to easily detectable DNA fragment polymorphism. Since polymorphic DNA fragments are derived from precise locations on the organism's genome, they can serve as reliable genetic markers, or landmarks to identify and locate genes.

A host of assays to detect DNA polymorphisms, and SNPs in particular, have been developed. In some of these assays (*e.g.*, RFLP [Botstein, D., White, R.L., Skolnich, M., Davis, R.W., *Am. J. Hum. Genet.* **32**:314-331 (1998)], CAPS [Konieczny, A. Ausubel, J.F., *Plant J.* **4**:403-410 (1993)], dCAPS [Neff, M.M.

Neff, J.D., Chory, J., Pepper, A.E., *The Plant Journal* **14**:387-392 (1998)], PIRA [Steinborn, R., Muller, M., Brem, G., *Biochim. Biophys. Acta* **1397**:295-304 (1998)]), polymorphic nucleotide sequences produce double strand cleavages either within or near the recognition sequence. The specificity of restriction enzymes is such that they exhibit a unique sensitivity to detect single nucleotide differences occurring in their recognition sites. The principal strengths of restriction enzyme-based genetic analyses are the east of use and the robustness of the assays. In the majority of the cases, the restriction site polymorphism is used to detect known, previously identified SNPs and the assay consists of any electrophoretical fragment analysis. In one report, the allelic variation is detected in a solid-phase ELISA-type setting [Truett, G.E., Walker, J.A., Wilson, J.B., Redmann, S.M. Jr., Tulley, R.T., Eckardt, G.R., Plastow, G., *Mamm. Genome* **9**:629-632 (1998)].

In WO 91/17269, Lerner *et al.* describe a different method for mapping a eukaryotic chromosome by restriction endonuclease mapping of discrete DNA sequences which are complementary to a region of a eukaryotic chromosome.

Vos *et al.*, *Nucl. Acids Res.* **23**:4407-4414 (1995) and EP 0 534 858 describe a technique for DNA fingerprinting called AFLP which is based on the selective polymerase chain reaction based application of restriction fragments of a digest of genomic DNA. The application reaction depends on the use of primers that extend into restriction fragments amplifying only those fragments in which prior extensions match the nucleotide sequence flanking the restriction sites.

Another method utilizing DNA amplification steps is set out in William *et al.*, *Nucl. Acids Res.* **18**:6531-6535 (1990) who describe a DNA fingerprinting method termed random amplified polymorphic DNA.

DNA amplification fingerprinting was described by Caetano Anolles in *Bio/Technology* **9**:553-557 (1991). Still another fingerprinting technique called arbitrarily primed PCR was described in Welsh *et al.*, *Nucl. Acids Res.* **18**:7213-7218 (1990) and Welsh *et al.*, *Nucl. Acids Res.* **19**:861-866 (1991).

In WO 94/11530, Cantor *et al.* describe materials and methods for position and sequencing by hybridization. Cantor *et al.* also describe methods for creating assays of DNA probes useful in the practice of their method.

The major shortcoming of the current methods of genetic analysis is the limited resolution of the DNA fragment analysis systems, namely the number of DNA fragments that can be separated in a single assay. Generally the fractionation resolution ranges from tens to a couple of hundred DNA fragments, at the most. Consequently, current genetic analysis methods are limited to a few hundred to a thousand genetic markers. While this resolution has been sufficient for analyzing simple genetic traits determined by single genes, the analysis of complex traits, which is now being undertaken and which involve general or many different genes, will require the analysis of a much larger number of genetic markers. It is anticipated that such studies will require from a few thousand to possibly several hundred thousand genetic markers. Although this could conceivably be accomplished by performing many parallel assays, such scaling up will be cost- and labor prohibitive.

A technology that has great potential and which is generating widespread interest in the so-called micro-array technology (DNA chips). In general, these methods are based on measurement of the hybridization of DNA sequences in solution to probe sequences that are arrayed on a solid surface. When assaying nucleotide polymorphisms, the detector relies on the small differences in hybridization efficiency between two different DNA sequences. In one format, fluorescently labeled sample DNA is hybridized to dense arrays of probe nucleic acids, sequence-specific hybridization signal is detected by scanning confocal microscopy, and DNA variants scored as (predictable) differences in the hybridization pattern. The micro-arrays are fabricated either by in-situ light-directed oligonucleotide synthesis [Fodor, S.P.A., *Science* 251:767 (1991)] or by spotting DNA (off-chip synthesized oligonucleotides or PCR fragments) in an automated procedure. The technology has already been demonstrated in the scoring of mutations in the mitochondrial [Chee, M., Yang, R., Hubbell,, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P.A., *Science* 274:610-614 (1996)] and HIV [Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N. *et al., Biotechniques* 19:442-447 (1995)] genomes as well as mutations in the CFTR cystic fibrosis gene [Cronin, M.T., Fucini, R.V., Kim, S.M., Masino, R.S., Wespi, R.M., Miyada, C.G., *Human Mut.*7:244-255 (1996)], the BRCA1 breast cancer gene (Hacia, G.H., Brody, L.C., Chee, M.S., Fodor, S.P.A., Collins, F.S., *Nat.*

Genet. **14**:441-447 (1996)], as well as for scoring random mutations in the yeast genome [Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., Davis, R.W., *Science* **281**:1194 (1998)]. In comparison with most other assays, micro-arrays provide a platform for high-throughput, massively parallel polymorphism detection.

A major disadvantage with the use of microarrays relates to the complexity of the hybridization reaction. The detection relies on the very small difference in hybridization of DNA sequences differing by only one nucleotide. In general, a set of 4 oligonucleotides, differing only in the identity of the central base, is synthesized for each position in the target sequence that has to be interrogated. The degree of redundancy further increases dramatically if one wants to screen the target DNA for all possible mutations; the design then includes overlapping oligonucleotide-sets that are offset by one base (a process known as tiling).

Another of the major drawbacks of the DNA chip technology is that each SNP marker must be PCR amplified individually from the sample DNA [Wang, D.G., *et al.*, *Science* **280**:1077-1082 (1998)]. Each high density SNP assay thus requires a number of different multiplex PCR reactions, each involving complex mixtures of PCR primers. It should also be noted that the detection of SNPs by hybridization to arrays depends on the use of short oligonucleotide probes. With longer probes such as DNA fragments in the size range of 50 to 500 base pairs or larger, it is not possible to distinguish the SNP alleles. While DNA microchips show great promise in the scoring of known SNPs, it remains to be demonstrated whether it will be an effective approach for large scale diagnosis of polymorphisms.

Detailed Description of the Invention

The methods of the present invention combine the robustness of DNA fragment analysis with the massive parallel measurement power of microarrays. The methods generally include the steps of:

(1) preparing sets of DNA fragments (probe DNA) which contain a particular kind of DNA sequence polymorphism and which are then used to prepare microarrays; and

(2) preparing concomitantly amplifiable sample DNAs such that thousands of sequence polymorphism's are detected by the presence or the absence of a hybridization signal on each of the probe fragments in the microarray after treatment with a probe enzyme and amplification.

Polymorphisms detectable according to the methods of the present invention include single nucleotide polymorphisms. In particular, the polymorphisms detected according to the present invention are those which give rise to a restriction endonuclease site or which eliminate a restriction endonuclease site. Such polymorphisms are referred to herein as endonuclease site polymorphisms (ESPs).

In one of its embodiments the present invention is directed to methods for detecting ESPs in a "restricted amplicon assay" (RAA) which comprises preparing concomitantly amplifiable target DNA (target DNA or amplicon DNA). The target DNA may be in the form of a restriction fragment of DNA with defined 5' and 3' termini. Target DNA fragments are typically prepared by digestion of DNA with a rare cutter restriction endonuclease (*i.e.*, hexacutter) and a frequent cutter (*i.e.*, tetracutter) collectively referred to herein as sampling enzymes or targeting restriction endonuclease reagents. The target DNA may be further modified at its termini by the addition of primers and/or adapters which may serve to prime an amplification reaction. Once target DNA is obtained, it is treated with a probing enzyme also referred to as a probe restriction endonuclease reagent which preferably has as a recognition site a nucleotide sequence of less than six nucleotides. More preferred probe restriction endonuclease reagents have a recognition site of four or fewer nucleotides. In certain embodiments, the probe restriction endonuclease reagent has a recognition sequence of two nucleotides which sequence is preferably (but not limited to) CpG. The probe restriction endonuclease reagent may comprise more than one restriction endonuclease so long as the size of its recognition sequence falls into one of the foregoing size ranges described for the reagents.

After treatment of the target DNA with probe restriction endonuclease reagent, the treated DNA is amplified, preferably by using a polymerase chain reaction. Subsequent to the amplification step, the target DNA is analyzed for the presence of an ESP by any of a variety of methods described below or by other methods well known in the art. In these methods, an ESP is identified either by the presence of a recognition site for the probe restriction endonuclease reagent (which will result in the failure of the target DNA to amplify) or by the loss of a recognition site which will allow amplification of an otherwise unamplifiable target DNA.

Arrays, or microarrays of probe DNA (as defined above) wherein the probe DNAs useful in the detection of ESPs are also encompassed by the present invention. Probe DNAs may be prepared by digestion of DNA with targeting enzymes as described above. Informative probe DNAs are then identified as described in detail below and are then attached to a substrate for use in the hybridization reactions with concomitantly amplifiable DNA after treatment with a probe restriction endonuclease reagent and subsequent amplification.

The present invention is also directed to methods for targeted restricted amplicon assays for the detection of ESPs. Targeted RAA operates on the same principal as random RAA except that the target amplicons need not be DNA fragments, but are rather defined amplifiable regions of a genome which are flanked by amplification adapters/primers. The amplicons of the targeted RAA may be identified using random RAA methods or by other methods such as direct sequence analysis of the DNA to be used as probe DNA. In targeted RAA, DNA to be analyzed is treated with a probe restriction endonuclease reagent, followed by the concomitant amplification of the treated DNA (amplicons) using predetermined primer using, for example, the polymerase chain reaction as described herein. The analysis of the amplification products then proceeds as described in the random RAA methods described herein. As with random RAA, an ESP is defined as the presence or absence of a recognition site for the probe restriction endonuclease reagent.

Since the method of the invention is based on the detection of a particular kind of DNA polymorphism which occurs in DNA of any organism, the invention will be universally applicable for genetic analysis. Furthermore, based on the large body of DNA sequence data at hand, it is predicted that the genomes of

higher organisms carry several hundreds of thousands of such DNA polymorphism. Consequently, the new method is capable of diagnosing the immense number of genetic markers that are needed to unravel complex traits. The method is of tremendous value for high throughput genetic analysis in the emerging field of pharmacogenetics. Similarly, the method has great potential in the field of animal and plant breeding, where high resolution genetic analysis will be needed to identify the genes involved in quantitative agronomic traits.

The following is a more detailed description of various aspects of the present invention. Variations in each of these aspects will be readily appreciated by one of ordinary skill in the art and one with the scope of the invention

I. Target DNA fragments (Amplicons)

(A) Fragment Size

The optimal fragment size for use in the methods (and materials) of the present invention is determined by two parameters: (1) size limits for synchronous amplification and (2) the optimal size for having on average one cleavage site for the probing enzyme.

It has been shown that random DNA fragments can be amplified synchronously when using a single PCR primer pair that attaches to the ends of the fragments. To this end synthetic oligonucleotides are ligated to the ends of restriction fragments. It has been observed that the synthetic sequences at the two ends of a fragment must be different, presumably because otherwise inverted repeats are generated at the ends that may form duplexes after denaturation. Consequently, a preferred mode for the method of the present invention involves the use of two different restriction enzymes (restriction endonuclease reagent): a rare cutter enzyme combined with a frequent cutter enzyme as described in EP 0 534 858 A1 which describes a method called AFLP and which is incorporated herein by reference (see figure 1) to prepare target DNAs (amplicons).

As can be seen from figure 1, the rare cutter enzyme produces large fragments that upon cleavage with the frequent cutter enzyme are cut into a number of smaller fragments. This dual cleavage generates two types of fragments: the majority having both ends produced by the frequent cutter (type I) and a minority of

fragments having a rare cutter end and a frequent cutter end (type II). After ligating different adapters to each of the ends and using appropriate primers targeted to the ends of the fragments, only the type II fragments will be amplified efficiently. Indeed the type I fragments carrying two frequent cutter ends will have inverted repeats at their ends and will amplify with greatly reduced efficiency. Consequently only the type II fragments are visualized in AFLP patterns.

In general all type II fragments will amplify synchronously up to a certain size limit. This size limit is dependent upon both the DNA polymerase used and the PCR reaction conditions. Under typical AFLP reaction conditions the size limit is around 500 bp. However, using different DNA polymerases, it is possible to increase this limit to 1000 bp or more.

The optimal size for obtaining an average of one cleavage per fragment with a probe restriction endonuclease reagent depends on the cleavage frequency of the reagent in the DNA under study. Hence, different probing enzymes will have different size optima.

Alternative schemes will be readily apparent to one of ordinary skill in the art from the one described above will perform equally well, such as the use of type III restriction enzymes having nucleotide recognition sequences or the use of pairs of frequent cutters.

(B) Fragment Complexity

The optimal complexity of the target fragment DNA sample is determined by two parameters: (1) the number of ESPs that are detected in a single assay, and (2) the detection sensitivity of micro-array hybridization.

In general the objective is to score as many ESPs as possible in a single assay. Hence the larger the number of starting fragments, the more ESPs can be scored. This number is however limited by the detection sensitivity of the assay. Based on published micro-array data the detection sensitivity is in the range of 1:50,000 to 1:300,000. In other words, one can detect one fragment in a fragment mixture with a complexity of 50,000 to 300,000 fragments. Assuming that all genomic fragments amplify with roughly equal efficiency, and assuming that all fragments containing ESPs occur in a single copy in the genome, then the detection limit is determined by the total number of genomic fragments that are amplified. In a

preferred embodiment of the present invention, target sample complexity should not exceed 100,000 fragments.

A close look at the procedure reveals that the total number of genomic fragments that will be amplified is determined by two components of the system: the sampling enzymes (target restriction endonuclease reagents) and the probe restriction endonuclease reagent (or probing enzymes).

Target restriction endonuclease reagents (sampling enzymes). As outlined above the number of amplifiable fragments will be determined primarily by the choice of the rare cutter restriction enzyme. In fact this number equals two times the number of target sites for the rare cutter. The number of target sites can in turn be determined by dividing the total genome size by the average size of the rare cutter fragments. In a preferred embodiment, restriction enzymes recognizing 6 nucleotides (hexacutters) or more are used as rare cutters in combination with frequent cutting restriction enzymes recognizing 4 nucleotides (tetracutters) or fewer.

Probe restriction endonuclease reagents (probe or probing enzymes) As probe restriction endonuclease reagents, different tetracutter enzymes can be used. Probing enzymes having recognition sequences of fewer than 4 nucleotides may also be used. Optimally these should cleave the target fragments only once on average: indeed if the probing enzyme cuts more than once, possible mutations affecting one of the recognition sites will remain undetected because the fragment will be cut at the non-mutated site. It should be realized that the cleavage with the probe restriction endonuclease reagent will have a considerable impact on the fragment complexity. Indeed, when the target fragments are cleaved once an average by the probing reagent, only about 35% to about 20% of them will be amplified. This means that the cleavage with the probing enzyme causes a 3 to 5-fold reduction in the fragment complexity.

In conclusion, the preferred complexity of 50,000 to 300,000 fragments in the final assay (see below) can be achieved by a judicious choice of sampling and probing enzymes. Preferably, a mixture of target fragments (amplicons) comprising 100,000 to 200,000 fragments is obtained and preferably a combination of a frequent cutter sampling enzyme and a probing enzyme are chosen such that 75% of

the target fragments are cut. This will produce a preferred complexity of 50,000 fragments.

II. Mutations Detected

In essence the method of the invention aims to detect mutations affecting the recognition sequences of the site-specific endonucleases which are used as probing enzymes. When the probe enzyme cleaves a target fragment, it is prevented from being amplified and as a consequence the target fragment will not give a hybridization signal with its cognate probe. Mutations affecting the recognition sequence of the probe enzyme will allow amplification of the target fragment and will restore the hybridization signal. A critical analysis of the entire process of the invention reveals that differences in hybridization signals may be due mutations other than those affecting the recognition sites of the probe enzymes. Indeed, any mutation that affects the target fragment or its amplification will lead to a loss of hybridization signal. In particular, mutations affecting the recognition sites of the sampling enzymes may also give the same result.

When using as sampling enzyme combinations of rare and frequent cutting restriction enzymes to generate target DNA fragments, one will generally obtain two target fragments flanking the rare cutter. Each of these fragments may carry recognition sites for probing enzymes that will affect their amplification. Consequently, the hybridization assay may detect two, one or none of the two fragments. The genetic variation in the germplasm consists mainly (90%) of point mutations. These can affect each of the recognition sequences as exemplified in figure 2.

Mutations affecting the probe enzyme sites (ESPs), only affect the amplification of the target fragment carrying the mutation (i.e., mutations giving rise to a probe enzyme recognition site will not amplify, those eliminating the recognition site will give rise to an amplifiable fragment). In contrast, mutations affecting the recognition sites of the sampling enzymes will have quite different consequences. Mutations affecting the rare cutter site will have as a consequence that the two target fragments will not be produced, and will thus prevent both fragments from being detected. Mutations affecting the frequent cutter site will in general have little or no

effect, because by the very nature of their frequent occurrence there is inevitably another frequent cutter site nearby which yield a larger fragment. Whether this fragment will or will not give a hybridization signal may depend on the presence of sites for the probing enzyme in the additional DNA segment.

In conclusion, the mere detection of a hybridization difference between two samples does not qualify the difference as being due to an ESP. For this one must also assay the two samples without probing enzyme cleavage, only those differences that are correlated with the cleavage by the probing enzyme qualify as genuine ESPs as defined according to the present invention.

In addition to the above issue two further points are worth commenting. The first is the issue of pseudo-alleles and second is the issue of haplotypes.

Pseudo-alleles. When performing genetic analyses of families, one follows the inheritance of traits relative to sets of genetic markers that differentiate the parents. In this case it should not matter whether one is typing bona fide ESPs or other types of mutations. In contrast, in population genetics a simple +/- assay cannot always distinguish between different types of mutations affecting the same fragment. In this case the positive (or negative) hybridization signal could be due to different mutations, and hence different alleles of a target fragment cannot be distinguished unless the same fragment can be analyzed with different probing enzymes. The combinatorial analysis now provides the possibility of distinguishing different alleles at the same locus.

Haplotypes. The term haplotype refers to particular combinations of genetic markers at a genetic locus. In general haplotypes rather than individual markers are used in population studies to detect statistical associations between traits and specific chromosome regions. Hence novel methods for genetic analysis should have the power to measure haplotypes. The simple realization that each rare cutter site generates two distinct target fragments already provides two building blocks for constructing haplotypes. Each of these can be dissected further using either different probing enzymes or even different frequent cutting enzymes.

III. Detecti n of ESPs

(A) Probe DNA

A feature of the present invention is that ESPs are detected by hybridizing sample DNAs to micro-arrays (or by other methods discussed below) comprising a set of probe DNAs which are designed such that each probe will hybridize specifically to one sample DNA fragment. For each set of target DNA fragments a specific set of probe DNAs are developed that will detect all the ESPs present in the set of target DNAs. Since in most applications only a minor fraction of the target DNAs will actually carry an ESP for a particular probe enzyme, the set of probe DNAs will consist of a subset of the target DNA fragments that are selected to carry ESPs. (It should be noted however that more elaborate strategies using multiplexing probe enzymes will eventually use most of the target fragments). Preferably, the ESP probes are highly specific for the target fragment carrying an ESP, and do not cross-hybridize with other fragments in the sample. This feature is verified by testing the candidate probe DNAs in hybridization assays using the sample DNAs.

It is important to stress that the method of the invention can be used with any type of micro-array: spotted ESP fragments, spotted ESP oligonucleotides or ESP oligonucleotides synthesized on solid supports using photolithography (Affymetrix). When using spotted fragments, these are obtained by cloning and amplifying target fragments. ESP oligonucleotides can easily be designed based on the nucleotide sequences of the ESP fragments.

The sections below describe different approaches that may be used to assemble sets of unique probe DNAs for fabricating the micro-arrays. Three alternative approaches are presented, and their choice is determined primarily by the degree of nucleotide sequence variation, and hence the ESP frequency, present in the samples under study.

Batch-wise hybridization selection method. Since both approaches described above are very inefficient and labor intensive when the ESP frequency is smaller than 5%, it is advantageous to directly select ESP fragments from the starting target fragments. Such an approach is described in greater detail below, and will be used in the human example (Example 3).

Direct screening. When the ESP frequency is high, such that 10% or more of the target fragments carry ESPs, the fastest approach for assembling ESP probe fragments is to array individual target fragments and test which of them detect an ESP in the samples under study. The advantage of this approach is that the same set of fragments can be tested with different probe enzymes. After the screening one will retain only those fragments that yield a clear-cut difference in hybridization between the different samples. This approach will be illustrated in the corn example (Example 2).

Gel-based screening. With samples exhibiting intermediate ESP frequencies (5% to 10%), one can use a gel-based screening approach in which the ESPs are identified by comparing the patterns of target fragments cleaved with the probing enzyme in different samples. The polymorphic fragments can then be isolated from the gel and cloned or amplified. These fragments still need to be verified in the micro-array hybridization assay. This approach will be illustrated in the Arabidopsis example (Example 1)

(B) Batch-wise hybridization selection method

The rationale for the positive selection for ESP fragments is to take advantage of the fact that a complementary procedure can be designed to select both for fragments that carry a probe enzyme site as well as for fragments that lack a probe enzyme site. By using a batch-wise hybridization, fragments can be selected that are present in common in both samples. For the sake of clarity we will term fragments carrying a site are terminal S+ fragments and fragments lacking the site S- fragments. In essence, the approach comprises four steps: (1) the preparation of sample DNA, (2) the preparation of S+ and S- fragments, (3) a hybridization selection step and (4) the amplification and isolation of ESP fragments. Each of these steps is described in detail below.

(i) Preparation of the starting DNA The preferred starting material is an equimolar mixture of genomic DNA from a number of individuals that is representative for the entire population of the species under study. Such a mixture can readily be obtained by mixing the same amount of genomic DNA from for example 20 to 50 individuals. After cleavage of the mixed DNA sample with combinations of sampling enzymes and ligation of the appropriate oligonucleotide

adapters as described above, a starting DNA fragment mixture is prepared. In this mixture of target fragments three classes of DNA fragments can be distinguished with respect to sites for the probing enzyme: fragments that always carry no site, fragments that always carry one or more sites and fragments that carry one site which is polymorphic. The latter class represents the fragments that carry ESPs and in the mixture of target fragments these will be present in the two forms (with and without the site, respectively the S+ and the S- fragments).

(a) Preparation of S+ and S- fragments The starting DNA is divided in two aliquots, which will be treated separately to prepare respectively the S+ fragment mix and the S- fragment mix:

S+ fragment mix The first step in this procedure is the amplification of the target DNA fragments using the standard procedure. Thereafter the amplified DNA is cleaved with the probing enzyme and appropriate oligonucleotide adapters (see EP 0 534 858 incorporated herein by reference) are ligated to the ends generated by the probe enzyme. By now setting up two polymerase chain reaction amplification reactions using one primer that recognizes the probe enzyme adapter and one primer that recognizes one of the two sampling enzyme adapters, one can specifically amplify those fragments that are cleaved by the probing enzyme. The two "halves" of these fragments will be amplified in either one of the two reactions. Furthermore by using biotinylated primers the resulting amplified S+ fragment mix products can be attached to solid substrates such as magnetic beads conjugated with streptavidin.

S- fragment mix To prepare the S- fragment mix, the target DNA fragments are cleaved with the probing enzyme and amplified. This will result in a mixture of fragments that do not contain sites for the probing enzyme.

(b) Hybridization selection step The S- fragment mix is hybridized to the S+ fragments bound to the magnetic beads. After extensive washing of the non-hybridized DNA, the annealed S- fragments are eluted and reamplified. Only the ESP carrying fragments will anneal to the fragments bound to the beads, since the other two classes of fragments are missing in one of the two fragment samples.

(c) Isolation of ESP fragments The candidate ESP fragments can now be isolated by cloning the amplified S- fragments and spotting them on micro-arrays for hybridization assays.

The following illustrative examples were chosen to represent the spectrum of genomic complexities and the spectrum of degrees of genetic variation which are susceptible to analysis using the methods of the present invention:

Example 1 describes analysis of Arabidopsis (low genomic complexity, low genetic variation).

Example 2 describes genetic analysis of corn (high genomic complexity, high genetic variation).

Example 3 describes genetic analysis in humans (high genomic complexity, low genetic variation).

It is also recognized that the methods of the present invention may be used to identify ESPs in a wide variety of organisms from procaryotic organisms, such as bacteria, through complex eukaryotic organisms, viruses, or any organism having a genome however simple or complex. The methods may also be used for the analysis of DNA libraries, for example, yeast artificial chromosome libraries and others.

Example 1

Genetic analysis in Arabidopsis

In this example, a fragment analysis-based approach (random ESP assay) is used to generate a set of genomic fragments carrying ESPs between the Arabidopsis ecotypes Landsberg and Columbia, which are commonly used for genetic studies in the model organism. The results described in this and the remaining examples is based on a computer-based analysis using publicly available DNA sequences (in silico analysis).

Arabidopsis is an example of a low complexity genome (size 100Mb), and the two ecotypes exhibit a moderate level of genetic variability. Extensive AFLP studies revealed that on average of 10% of the fragments are polymorph between the two ecotypes. This corresponds to a difference of 1 in 150 nucleotides. Consequently, the fraction of fragments expected to carry an ESP for tetranucleotide recognizing

restriction enzymes is expected to be in the range of 2.5% (1:40). With such a low frequency, it is helpful to use a selection procedure to isolate the rare fragments containing ESPs.

In essence the procedure described in this example comprises the following steps:

1) Identification of a set of about 200 genomic fragments carrying Landsberg / Columbia ESPs using a gel-electrophoretic approach.
2) Isolation and characterization of the ESP carrying DNA fragments (ESP fragments).
3) Generation of micro-arrays with the ESP fragments and confirmation of the ESPs by hybridization of a mixture of differentially labeled Landsberg and Columbia genomic DNA.

Step 1. Identification of ESP fragments.

Sampling enzymes. In the present example EcoRI, a restriction enzyme recognizing 6 nucleotides (hexacutter), in combination with BfaI, a restriction enzyme recognizing 4 nucleotides (tetracutter), are chosen as sampling enzymes. From the random frequency of occurrence of 6 nucleotide sequences (every 4,000 bases), the number of sites for hexacutter restriction enzymes in this genome is predicted to be in the range of 25,000. In addition to cleavage with a hexacutter, the genomic DNA is also cut with a tetracutter so as to generate PCR amplifiable fragments of an average size of a few hundred base pair. Since the majority of the hexacutter fragments will give rise to two fragments having a rare cutter end and a frequent cutter end (see figure 1), this procedure will yield a mixture of about 50,000 fragments.

Probing enzymes. As probing enzymes many different tetracutter enzymes can be used. Ideally, the probing enzyme cleaves most of the sample fragments once. Because plant DNA has a high AT content, the preferred tetracutters are those that have an AT bias in their recognition sequence. In general, the choice of an optimal tetracutter may be determined by particular features of the genome being analyzed (*e.g.*, AT and GC content) In the present example, MseI (recognition site = TTAA) was chosen. Tsp509I (recognition site = AATT) is an alternative. It is also

conceivable to use mixtures of two or more tetracutter enzymes. The EcoRI-BfaI sample/target fragments that are cleaved and not cleaved with the MseI probing enzyme are referred to as cleaved and uncleaved sample/target DNA, respectively.

Screening for ESP carrying fragments. To detect ESP fragments, subsets of uncleaved and cleaved EcoRI-BfaI sample fragments from both ecotypes are amplified and the amplicons are compared following gel-electrophoretic fractionation. The AFLP-method to selectively amplify specific subsets of the EcoRI-BfaI sample fragments is used [Vos, P. *et al., Nucleic Acids Res.* 23:4407-4414 (1995); Zabeau, M. and Vos, P., European Patent Application EP 0534858 (1993) both of which are incorporated herein by reference]. Given the complexity of the sample (~50,000 fragments), the selective amplifications are performed with EcoRI and BfaI primers having two and three selective nucleotides, respectively. This equals 1024 (16 x 64) different selective amplification reactions.

The experimental procedure described by Vos P. *et al.* is followed except that the template fragments are purified and, when applicable, digested with the probing enzyme prior to amplification. The structures of the EcoRI and BfaI adaptors are as follows [see, *e.g.*, Vos, P. *et al., supra*]:

```
5'-CTCGTAGACTGCGTACC
      CATCTGACGCATGGTTAA-5'

5'-GACGATGAGTCCTGAG
       TACTCAGGACTCAT-5'
```

The EcoRI (radiolabeled by 5'-phosphorylation) and BfaI primers, having two and three selective nucleotides, respectively, have the following sequences (where N represents A, C, G, or T):

```
5'-GACTGCGTACCAATTCNN

5'-GATGAGTCCTGAGTAGNNN
```

Using these reagents, most of the obtainable target fragments contain a cleavage site for the probing enzyme and, consequently, disappear when the target

DNA is cleaved. Most of the fragments that survive the treatment with the probing enzyme occur in both ecotypes, and thus carry no ESP. Occasionally fragments are found that appear in both ecotypes when the target DNA is not digested and that are present in only one of the two ecotypes after digestion. These represent true ESPs for the probing enzyme. The fragments will also show typical AFLP polymorphism between the two ecotypes. Such polymorphisms are apparent in the fragment patterns obtainable with the undigested sample DNAs. Again, the majority of the polymorphic AFLP fragments will not be present in the samples treated with the probing enzyme.

Systematic comparison of the patterns of ecotypes Columbia and Landsberg before and after digestion, allows the identification of EcoRI-BfaI sample amplicons that carry an ESP for the probing enzyme. Using MseI as sampling enzyme, we have identified a total of ~200 polymorphic fragments which are present in only one of the ecotypes.

Step 2. Isolation and characterization of ESP fragments.

Each of the ESP polymorphic fragments is eluted from the gel-matrix, re-amplified using EcoRI and BfaI AFLP primers with no selective nucleotides and cloned into a suitable plasmid vector (e.g. TA cloning system; Invitrogen, Carlsbad, CA, U.S.A.). In each case, two clones are selected for sequence determination. Most duplicate clones will yield the same sequence. Duplicate clones that gave different sequences were not retained for further work. Since the nucleotide sequence of over one third of the Arabidopsis genome is available in the public databases (*e.g.*, Genbank), the chromosomal location of one third of the ESP fragments can be determined by matching the fragment sequences to the genomic sequence. Furthermore since the genomic sequence is derived from ecotype Columbia, we expect a perfect match with the fragment sequences isolated from the same ecotype is expected. The sequences of the fragments isolated from ecotype Landsberg will reveal single nucleotide differences, amongst which the potential restriction site mutations should be apparent.

Step 3. Fabrication of ESP micro-arrays.

Micro-arrays of amplified fragments. The insert DNAs from the sequence verified clones are amplified using the non-selective AFLP primers. PCR products are verified by agarose gel electrophoresis and retained if a single product of the correct mobility was present. Following ethanol precipitation, the resuspended PCR products are arrayed at high density on standard glass slides (25 x 76 mm) using either the Multigrid robotic spotter (Gene Machines, Menlo Park, CA, U.S.A.) or the BioChip Arrayer™ (Packard Instrument Company, Meriden, CT, U.S.A.) (μL/nL per spot). The DNAs are spotted in a logical order with respect to the probing enzyme used (left and right panels) and the ecotype from which the fragments were isolated (upper and lower panel) as shown in figure 3.

Micro-arrays of oligonucleotides. Based on the nucleotide sequences of the ESP fragments, oligonucleotides can be designed that can serve as hybridization probes to specifically detect each amplified sample fragment. The oligonucleotide probe should preferably match with a sequence that is located to one side of the ESP, opposite the side where the sequence targeted by the labeled primer is located. In this way the background is minimized because the linear amplification products generated by the labeled primer following digestion with the probing enzyme are not detected. The ESP fragment specific oligonucleotides are spotted in a micro-array format in exactly the same way as the amplified ESP fragments.

Step 4. Micro-array-based detection of ESPs.

Preparation of the sample DNAs For each ecotype, sample DNA is prepared in two different ways. Genomic DNA, digested with the sampling restriction enzymes EcoRI and BfaI, was amplified either as such or after cleavage with the probing enzyme MseI. The amplification reactions are performed with a fluorescently labeled EcoRI primer and an unlabeled BfaI primer, both without selective (AFLP) nucleotides. For preparation of the Columbia samples a Cy3(green)-labeled EcoRI primer is used, whereas the Landsberg-derived fragments are amplified with a Cy5(red)-labeled EcoRI primer. Cy3- and Cy5-amidites are incorporated during primer synthesis (Amersham Pharmacia Biotech, Uppsala, Sweden). Two different hybridization solutions are then prepared, one by mixing equal amounts of the

uncleaved sample of both ecotypes, and a second containing the two cleaved samples. These mixtures are obtained by mixing equal volumes of the respective amplification reactions; the overall complexity of the mixed samples may be assumed to be identical, while also the PCR reaction conditions were the same for the two ecotypes. It should be realized that the probing enzyme cleaves most of the sample fragments and that, therefore, the complexity of the two sets of sample DNA, cleaved versus uncleaved, differs considerably. It is estimated that the uncleaved sample contains roughly 4 times as many fragments.

In case arrays of PCR products, rather than oligonucleotides, are used as probes (refer to step 3), the co-amplification of the EcoRI-BfaI sample fragments is preferably accomplished with a pair of adaptors that differs from those attached to the arrayed probes. The alternative EcoRI and BfaI adaptors have the following structure:

```
5'-GAGCATCTGACGCATCC
      GTAGACTGCGTAGGTTAA-5'

5'-CTGCTACTCAGGACTG
       ATGAGTCCTGACAT-5'
```

The cognate non-selective EcoRI and BfaI primers have the following sequences:

```
5'-CTGACGCATCCAATTC

5'-CTACTCAGGACTGTAG
```

Micro-array hybridization. Each of the hybridization solutions is allowed to hybridize to the arrayed probes using protocols well known in the art. The experimental conditions depend primarily on the nature of the probes, PCR-amplified fragments versus oligonucleotides. Both types of experiments are amply described in literature: Wodicka, L. *et al., Nature Biotechnol.* **15**: 1359-1367 (1997); Lockhart, D. J. *et al., Nature Biotechnol.* **14**: 1675-1680 (1996); DeRisi, J. L. *et al., Science* **278**: 680-686 (1997); Shalon, D. *et al., Genome Res.* **6**: 639-645 (1996); Piétu, G. *et al., Genome Res.* **6**: 492-503 (1196); Chee, M. *et al., Science* **274**: 610-614 (1996); Wang

D.G. *et al., Science* 280: 1077-1082 (1998); Winzeler E. A. *et al., Science* **281**: 1194-1197 (1998), all of which are incorporated herein by reference.

A laser scanning system (ScanArray 3000; General Scanning Inc., Watertown, MA, U.S.A.) is used to detect the two-color fluorescence hybridization signals from the micro-arrays at a resolution of 10 micron per pixel. A separate scan is carried out for each of the two fluorophores used. Scanning parameters and laser power settings are adjusted to normalize the signal in the two channels (channel-1/Cy3; channel-2/Cy5). The obtained digital images were analyzed using the ImaGene™ image analysis software (BioDiscovery Inc., Los Angeles, CA, U.S.A.). The extracted quantitative data are transferred to a spreadsheet for further analysis.

The present hybridization experiment is essentially set up as a confirmation of the gel-electrophoretic data (refer to step 1), and has, therefore, a predictable outcome. In addition, a number of control probes are included on the biochip that detect monomorphic EcoRI-BfaI Arabidopsis fragments (i.e., fragments on which a site for the probing enzyme is either present or absent in both ecotypes. Taken together, the results allow correction for background and optical cross-talk between the two channels, as well as calibration of the red and green hybridization signals. It is anticipated that the vast majority of the processed data are unambiguous with respect to the allelic state of a sample fragment and in agreement with the gel-electrophoretic analysis. Figure 4 shows a two-tone representation of a false- color display of the idealized results of the present experiment using a fictitious array of probes. In figure 4, yellow is represented by the light circles and green by the dark circles. It cannot be excluded that certain hybridization results are not in agreement with the gel-electrophoretic assay and/or that certain probes do not allow unambiguous determination of the allelic state of the cognate sample fragment. Such probes should be excluded from the micro-arrays that are used to genotype experimental Arabidopsis samples, other than the Columbia and Landsberg controls used in the present example.

An important feature of the ESP-method is that the complexity of the sample is reduced considerably by cleavage with the probing enzyme prior to amplific/ation. The number of Arabidopsis fragments generated by the EcoRI and BfaI sampling enzymes (~50,000) is estimated to be reduced about 4-fold by

digestion with the MseI probing enzyme. The reduced sample complexity 'facilitates' the hybridization reaction. In the present illustrative example, a hybridization experiment with a mixture consisting of uncleaved sample DNA of both ecotypes is included. In a routine genotyping experiment such uncleaved sample DNA would normally not be included and allele-calling would only involve a comparison of the signals obtained with cleaved test-sample and an appropriate control (e.g. cleaved Columbia or Landsberg sample DNA). The test-sample and control can, in principle, be hybridized separately but a preferred method consists of hybridizing a mixture of differentially labeled test- and control-sample to the same array. Inclusion of the control should allow to determine the zygosity following the various corrections and normalization procedures (refer to Example 2, step 3).

Example 2

Genetic analysis in corn

In this example, the utility of the method of the invention for marker assisted selection applications in plant and animal breeding is illustrated. Corn has been chosen because it is a typical representative of crop species having a complex genome. The large size of the genome (2,500 Mb), the frequent occurrence of repetitive DNA sequences and the high degree of genetic variation, all constitute technical challenges. In this example, an approach based on the generation of a set of genomic fragments carrying ESPs from two well-known inbred lines of corn, B73 and Mo17 from which many of the corn elite lines are derived is used. Another reason for choosing these lines is that a well-studied recombinant inbred population derived from these lines is available. This population can be used to map the set of ESPs. The genetic map of ESP markers will prove to be an effective tool for genetic selection in corn breeding. It is evident, however, that a broader survey of the corn germplasm with a total of 10 to 20 lines will give a much higher yield of ESPs (possibly 2 or 3 times as many) and will eventually result in a higher-resolution genetic map.

The ESP-harboring fragments could very well be identified by the gel-electrophoretic approach described for Arabidopsis (Example 1). However, an alternative strategy may be used given that the corn germplasm, like many crop species, exhibits a high degree of genetic variation. Indeed, based on data from AFLP

studies, the average nucleotide sequence variation in the corn germplasm is estimated to be in the order of 1 difference in 15 to 30 nucleotides. This corresponds to a frequency in ESPs in the recognition sites of tetracutter restriction enzymes of 1 in 4. At this frequency it becomes feasible to directly examine arrays of random B73/Mo17-fragments for the presence of ESPs making use of the present 'Restricted Amplicon Assay', without prior screening or selection. The strategy also lends itself readily to screening with two different probing enzymes, offering the opportunity to distinguish more than two alleles at certain loci.

In the present example, an approach is used in which individual ESPs are selectively amplified rather than sampling the DNA with a pair of restriction enzymes. This method is referred to as a targeted ESP assay. For each ESP one designs two specific PCR primers flanking the restriction site. As in the original approach, if the site is intact, the genomic DNA will be cleaved and no PCR product will be generated. Only when the site is mutated will the PCR product be generated. A similar approach whereby a large collection of specific primer sets are used to sample a test DNA was recently described for interrogating the allelic state at human SNP sites [Wang, D. G., *Science* **280**:1077-1082 (1998) incorporated herein by reference].

In essence the procedure described in this example comprises the following steps:

(1) Identification of a set of candidate ESP fragments from the inbred lines B73 and Mo17

(2) Development of a corn ESP micro-array

(3) Genetic mapping of a B73/Mo17 recombinant inbred population and of segregating populations

Step 1. Identification of candidate ESP fragments

Cloning of a set of target fragments. To clone a set of random fragments from the inbred lines B73 and Mo17, the enzyme combination SseI and BfaI is used. The octanucleotide-recognizing enzyme SseI was chosen because of the large size of the corn genome. It is estimated that this enzyme has around 5,000 to 10,000 sites in the corn genome. The second tetracutter-enzyme, BfaI, is expected to cleave in the majority of the cases on both sides of the SseI sites. The double

digestion will therefore generate between 10,000 and 20,000 target fragments with an average size of 400-500 base pair.

Following double digestion of the genomic DNA, SseI- and BfaI-adaptors were ligated to the fragment ends and the material amplified with non-selective SseI and BfaI primers. The structures of the SseI- and BfaI-adaptors are based on those described by Vos P. *et al., Nucleic Acids Res.* **23**:4407-4414 (1995):

```
5'-CTCGTAGACTGCGTACATGCA
   3'-CATCTGACGCATGT

5'-GACGATGAGTCCTGAG
     3'-TACTCAGGACTCAT
```

The corresponding SseI and BfaI non-selective primers have the following sequences:

```
5'-GACTGCGTACATGCAG

5'-GATGAGTCCTGAGTAG
```

The amplification step enriches the SseI-BfaI fragments over the large excess of BfaI-BfaI fragments. After amplification the fragments are fractionated on an agarose gel to eliminate the fragments smaller than 100 base pair, and cloned in an appropriate vector (e.g. TA cloning system; Invitrogen, Carlsbad, CA, U.S.A.).

Preparation of spotted micro-arrays with the cloned target DNA fragments. The insert DNAs, from the two libraries of cloned SseI-BfaI target fragments (obtained from the B73 and Mo17 inbred lines), are amplified from the clones using the non-selective SseI and BfaI primers. Following purification and concentration, the amplicons are arrayed as described in example 1. A total of 20,000 (i.e. 10,000 from each library) candidate probe DNAs are spotted.

Micro-array hybridization and selection of candidate ESP-fragments. From genomic DNA of the inbred lines B73 and Mo17 three different sets of

SseI/BfaI-digested amplified DNA are prepared. An alternative pair of adaptors and non-selective amplification primers are used for this:

5'-GAGCATCTGACGCATGTTGCA
3'-GTAGACTGCGTACA

5'-CTGCTACTCAGGACTG
3'-ATGAGTCCTGACAT

5'-CTGACGCATGTTGCAG

5'-CTACTCAGGACTGTAG

The target DNA is amplified either as such or after digestion with one of two alternative probing enzymes, MseI and Tsp509I. As probing enzymes many different tetracutter enzymes can be used. Because plant DNA has a high AT content, the preferred tetracutters are those that have an AT bias in their recognition sequence such as MseI and Tsp509I. Alternatively, mixtures of two or more tetracutter enzymes can be used.

For each of the B73 samples, a Cy3(green)-labeled SseI primer is used, whereas the Mo17-derived fragments are amplified with a Cy5(red)-labeled SseI primer (refer to Example 1). Different hybridization solutions are then prepared by mixing equal amounts of the uncleaved, MseI-cleaved, and Tsp509I-cleaved samples of both inbred lines. Each of the 3 mixes is allowed to hybridize to the micro-arrays. Analysis of the scanned images involved normalization using the multitude of probes on the arrays that detect monomorphic fragments.

Analysis reveals that candidate ESP fragments are readily identified by scoring differences in the hybridization images obtained with the two inbred line sample DNAs after cleavage with the probe enzyme. The quantitative analysis allows us the use of an unambiguous cut-off threshold of 10-fold difference for scoring ESPs. It should be pointed out that the polymorphisms identified in this assay may result from both bona fide ESPs for the probe enzyme and ESPs for the sampling enzymes. Part of the latter polymorphisms are eliminated by verifying the differential hybridization obtained with the sample DNAs not cleaved with the probe enzyme.

Analysis of 80 probes reveals that roughly 10% of the target fragments carry ESPs for either MseI or Tsp509I, in accordance with the expected ESP mutation frequency.

Probes exhibiting the hybridization pattern shown in the Table here below are retained for further analysis. Only fragments that do carry a recognition site for the probing enzyme are retained.

	B73/Mo17 (Cy3/Cy5) normalized hybridization signal		
	Undigested	MseI-digested	Tsp509I-digested
B73-probes	~1	< 0.1	< 0.1
Mo17-probes	~1	> 10	> 10

Step 2. Development of a corn ESP micro-array

Sequencing of the candidate ESPs and design of marker specific primers. Clones corresponding to the probes that yield the desired hybridization pattern (see Table) are sequenced. The majority of the insert DNAs derived from these clones will contain a single recognition site for the probing enzyme. For each unique candidate ESP, two specific PCR primers, flanking the restriction site, are designed.

In addition, the sequence of a limited set of probes that yielded invariant hybridization signals is also determined. PCR primers targeting these monomorphic sequences are included as references; they are used to calibrate the hybridization signals.

Validation of the candidate ESPs and fabrication of corn micro-arrays. The candidate ESPs, identified under step 1, are subjected to a confirmatory experiment using the marker specific primers. The experimental set up, however, differs considerably. First, the sampling is now performed by a set of specific PCR reactions rather than by a single co-amplification reaction of the SseI-BfaI fragments. Particular sets of the ESP-specific primers are combined in a series of multiplex PCR reactions; these reactions were in turn pooled to obtain the final set of sampling amplicons [Wang, D. G., *Science* **280**:1077-1082 (1998)]. Second, the hybridization mixtures are assembled in a different way. One of the ESP-specific primers is either Cy3- or Cy5-labeled; the other remained unlabeled. The Cy3-primer is used for amplification of the sample DNA that had previously been digested with the probing

enzyme, whereas the Cy5-primer is used in the case of undigested control sample DNA. A hybridization mixture consists of equivalent amounts of the two different sample preparations. The B73 and Mo17 derived material is analyzed in parallel. Third, the set of ESP-specific unlabeled primers also serves as hybridization probes. Arraying of these oligonucleotides is done in the same way as for amplification products. Fourth, the appropriate conditions are used for hybridization against oligonucleotide probes and are readily determined by one of ordinary skill in the art.

Direct comparison of the normalized Cy3 and Cy5 hybridization signals allows determination of the allelic state of the endonuclease target site in B73 versus Mo17. Primer pairs that do not allow unambiguous allele calling or that do not confirm the candidate ESPs identified with SseI-BfaI sampling (refer to step 1), are not retained for further work.

It is anticipated that each probing enzyme will identify roughly 2,000 ESPs that can be unambiguously scored in routine micro-array assays. When performing the above screening procedure (step 1 and 2) with a set of sample DNAs derived from 10 to 20 well-chosen commercial inbred lines, one may expect to find as many as 3,000 to 5,000 ESPs for each probe enzyme. The number may be further increased considerably by using sampling enzymes that yield more target fragments (such as the use of the hexacutter PstI instead of the octacutter SseI).

Step 3. Genetic analysis of a B73/Mo17 recombinant inbred population and of segregating populations

Genetic analysis of a B73/Mo17 inbred population. A collection of recombinant inbred lines derived from a cross between B73 and Mo17 is publicly available and provides a most useful set of lines for verifying and mapping the collection of ESP markers. The advantage of recombinant inbred lines over segregating populations is that each inbred line contains a different set of homozygous chromosome segments derived from either parent line. Consequently each ESP will be scored as either present or absent. Preparation of the sample DNAs and hybridization against the arrayed probes are performed as described under step 2. The experiment will, in the first place, allow the testing of selected ESPs in over 100 measurements; the results will result in the development of a second generation

system that will only detect the most consistent ESPs. In addition, the linkage analysis of the segregation data will allow the construction of a fine genetic map of the markers (using mapping data from other DNA markers such as RFLP's and AFLP's). Finally, based on the mapping data, an ordered ESP micro-array is developed for corn.

Genetic analysis of segregating populations. While isolated from two inbred lines, it is anticipated that the above-mentioned ordered ESP micro-arrays will detect sufficient genetic polymorphism in other corn lines to be useful for marker assisted selection. To demonstrate the applicability, one could either chose a segregating F2 population or a back-cross population. Sample preparations and hybridizations are again performed as described under step 2. In this experiment, the ESP markers must be scored quantitatively so as to differentiate between heterozygosity and homozygosity. Because only the most consistent markers are retained, a two-fold difference in signal intensity is easily monitored. The approach used consists of normalizing the hybridization signal intensities and then apply a mixture model analysis on the normalized data. This statistical approach consists of determining whether the relative signal intensities can be grouped into three discrete classes, corresponding to respectively homozygous present, heterozygous and homozygous absent. ESP markers that do not fulfill this criterion should be eliminated from the analysis.

Example 3

Human genetic analysis

This example illustrates the application of the method of the invention for genome-wide genetic analysis. Human is an example of a high complexity genome (size ~3000 Mb) combined with a very low level of genetic variability. Single nucleotide differences between pairs of allelic sequences from different individuals occur approximately once in every kilobase and estimations for the frequencies of ESPs in the population at large range from 1:500 to 1:200. As with Arabidopsis, such a low frequency necessitates the use of a selection procedure for the isolation/enrichment of the rare ESP-harboring fragments. In this example a batch-wise hybridization, such as that described in Example 1, is used to accomplish this.

Based on the known mutation frequencies, it can be estimated that the ESP frequency for a tetracutter-probing enzyme is in the order of 1 in 50 to 1 in 125. If the estimate of the micro-array detection limit of 1:100,000 is correct then the maximal number of ESPs that can be detected in a single assay is in the order of 800 to 2,000, assuming that all target fragments have one site for the probing enzyme. In reality the actual number will be only 200 to 500 (see below).

However, these limitations are overcome by taking advantage of a special class of probing enzymes to obtain ESPs at much higher frequencies. Indeed it is a well documented fact that a substantial fraction (>50%) of the nucleotide substitutions in the human genome result from C → T transitions in CpG dinucleotides. Such CpG dinucleotides represent mutational hotspots in vertebrates because a large fraction of the cytosines are methylated and subsequently mutate to thymine by deamination. It is estimated that the mutation frequency of methylated cytosines is 6 to 8-fold higher than average. Hence probing enzymes that recognize CpG will yield ESPs at correspondingly higher frequencies, estimated at 5% to 10%. However, the adverse consequence of the high mutation rate is that CpG is relatively rare in mammalian DNA, occurring with a frequency of 1 in 100 instead of 1 in 16. Likewise the cleavage frequency of CpG recognizing tetranucleotide restriction enzymes is 1 in 2000 instead of 1 in 256 bases. To compensate for this, a probe restriction endonuclease reagent comprising a cocktail of complementary restriction enzymes can be used; e.g. TaqI (TCGA), MspI (CCGG), MaeII (ACGT), and HinPI or HhaI (GCGC). It should be noted however that cleavage by the isoschizomers HinPI and HhaI is blocked by methylation of the cytosine residue (C^5) within the CpG dinucleotide. These enzymes will thus only cleave at a fraction of their sites, namely the non-methylated sites. The analysis of the large amount of human genomic DNA sequence shows that the cocktail of the 4 CpG enzymes will cleave once in every 400 bp on average. The total number of sites in the genome is thus in the order of 7.5 million. Assuming that the ESP frequency is 5% to 10%, the cocktail of the 4 CpG enzymes has the potential of detecting 375,000 to 750,000 ESPs.

Alternatively, the assay may use the endonuclease Cvi JI which specifically recognizes CpG (see WO 94/21633 incorporated herein by reference). The clear advantages are that the cleavage frequency of this enzyme is 4-fold higher

than that of the CpG enzyme cocktail and the potential spectrum of mutations that can be scanned is even 4 times larger.

Contrary to the previous examples where various combinations of sampling and probing enzymes may be used, the probing enzyme in the human application should be either one of the CpG enzymes or a cocktail of CpG enzymes. Consequently, the frequent cutter-sampling enzyme must be carefully chosen to cleave with a compatible frequency. Long stretches of human genomic sequence available in databases such as Genbank are used to investigate which sampling-probing enzyme combination yields the largest number of ESPs per assay.

In essence the *in silico* analysis (computer analysis) consisted of determining how many probing sites could be monitored in a single assay by calculating the frequency of target fragments carrying a single CpG enzyme site. The following assumptions were made: (1) the total target fragment complexity is 120,000 (2) only fragments larger than 50 bp and smaller than 1000 bp can be monitored. The results revealed that depending on the sampling enzyme used only 20% to 30% of the CpG sites can be monitored for ESPs. With an estimated 5% ESP frequency and using the cocktail of CpG enzymes the maximal number of detectable ESPs was 1,250. The main reason for this low number is that around 60% of the CpG sites occur in clusters. Similar results are obtainable using simulations with randomly generated cleavage sites. Thus, the silico analysis demonstrates that simple mathematical calculations using cleavage and mutation frequencies can be far away from reality. The conclusion for this analysis is that the development of a competitive technique using the method of the invention requires a radically different approach.

The principal reason why an approach based on random fragment sampling fails to yield a good number of ESPs in human DNA is that only a very small fraction (1% to 2%) of the target fragments carry a potential ESP. Because the number of target fragments is limited by the sensitivity of the hybridization assay, the total number of detectable ESPs per assay is limited to 1,000 to 2,000 at most. Since the low output is a direct consequence of the random nature of the fragment sampling strategy, the solution is to use a non-random target sampling: namely an approach in which individual ESPs are selectively amplified. In fact the design of such an approach is very simple. For each ESP one designs two PCR primers flanking the

restriction site and the genomic DNA is amplified after cleavage with the probing enzyme. Like in the original approach, if the site is intact, the DNA will be cleaved and no PCR product will be generated. Only when the site is mutated will the PCR product be generated. The terms "**random ESP assay**" and "**targetted ESP assay**" (as described above) are used to distinguish between the two approaches. The approach is perfectly feasible, as is evident from the recent paper by Wang *et al.*, *Science* **250**:1077-1081 (1998), incorporated herein by reference, in which it is demonstrated that it is possible to multiplex amplify 2000 SNP's in a limited number of PCR reactions.

The approach described here is in fact identical in the way the ESPs are amplified, but fundamentally different in the way they are diagnosed. The present method takes advantage of the clear distinction between having or not having a PCR product depending upon the allelic state of the endonuclease target site. The Wang *et al.* approach in contrast relies on the detection of a difference in a single nucleotide residue in the PCR product. This requires a much more elaborate and redundant assay. Finally, targeted ESP assay constitutes a nice extension of the method of the invention, namely the use of restriction enzyme digestion to detect SNP's. Furthermore the method also provides a means to monitor mutations in specific genes or loci instead of scanning the entire genome. Indeed, sets of ESP primers that are derived from a specific gene or chromosome region can be assembled.

In an alternative embodiment, the random ESP assay is combined with an assay in which sets of target fragments are subject to ESP analysis, followed by the detection of individual ESP fragments using fragment-specific PCR primer pairs. If the endonuclease treatment abolishes the amplification of the target fragment in the first round, then the specific primers will not give a PCR product. In this way the PCR primers must not flank the restriction site but can be directed to any part of the target fragment. Another potential advantage of this combined approach is a more synchronous first amplification round allowing all fragments to be amplified to the same extent. The second amplification round then comprises only a limited number of PCR cycles, only serving the purpose of generating a detectable amount of the fragment-specific PCR product. In this way, it is possible to obtain a more quantitative assay in which heterozygous and homozygous ESPs can be distinguished.

In summary, the ESP of the present invention provides the ability to overcome the difficulties resulting from the very low frequency of genetic variation. This approach will allow the development of human genetic assays that can monitor in the order of 2,000 to 5,000 ESPs or more.

The example described below illustrates the approach in a limited scale assay which characterizes the human ESPs for CpG enzymes that can be detected using the sampling enzyme combination PacI – BfaI. The rare cutter PacI was chosen because it is one of the enzymes that cleave with the lowest frequency (60,000 sites) in the human genome. In this the assay can start from a moderate complexity of 120,000 target fragments. BfaI was chosen because it generates fragments of an average size of 340 bp, large enough to capture a sizable number of CpG restriction sites, estimated in the order of 40,000. Assuming a 5% to 10% ESP frequency, the number of detectable ESPs is in the order of 2000 to 4000. It should be stressed that many different sampling enzyme combinations can be used and that thus a substantial fraction of the 375,000 to 750,000 CpG-enzyme ESPs can be monitored using the approach methods of the present invention.

In essence the procedure described in this example comprises the following steps:

(1) Development of a set of PacI - BfaI ESP probe fragments.

(2) Development of the targeted ESP assay

(3) Genetic analysis of human individuals

Preparation of PacI - BfaI ESP probe fragments

In this method the procedure of batch-wise hybridization for generating CpG enzyme ESP probe fragments based on the PacI - BfaI pair of sampling enzymes is used. The procedure is in essence as described in Example 1 and comprises the preparation of two complementary sets of amplified genomic fragments (S+ and S- fragments). S+ fragments are obtained by first amplifying the mixture of PacI – BfaI fragments, digesting these with one of the four CpG enzymes, ligating adapters to the CpG ends and then amplifying the fragments using a biotinylated CpG primer and either one of the PacI or BfaI primers. The S+ fragments are then bound to magnetic beads coated with streptavidin. S- fragments are obtained by amplifying the PacI – BfaI fragments after digestion with one of the CpG enzymes. The genomic

DNA used in this procedure is a mixture of DNAs from individuals from different ethnic groups obtained by mixing equal amounts of DNA. After hybridization of the S- fragments to the S+ fragments bound to the beads, the beads are repeatedly washed to remove all unhybridized fragments and thereafter the hybridized S- fragments are eluted. These are then reamplified with the PacI or BfaI primers and the hybrid selection procedure is repeated at least once. Finally the amplified fragments are cloned in an appropriate vector and a series of around 2,000 fragments are sequenced. To select a set of S+ fragments, this procedure is repeated in reverse using this time a biotinylated PacI or BfaI primer and binding the S- fragments to the beads. Upon comparison of the S+ and S- fragment sequences ESP fragments are readily identified as fragments having partially overlapping sequences and in which the S- fragment sequence shows a mutated CpG enzyme site at the internal boundary of the overlap. In this way we will characterize 500 to 1000 ESPs for each of the CpG probing enzymes.

Targeted ESP assay

This step comprises the following:

a) Design of ESP specific primers. For this, a procedure similar to the one described by Wang *et al.* to design PCR primers flanking the ESP site.

b) Preparation of ESP fragment probes. The procedure described earlier is used to prepare amplified DNA from the cloned ESP fragments for spotting on the micro-arrays. Alternatively, oligonucleotide probes can be designed to hybridize the ESP specific PCR products and spot these on the micro-arrays.

c) Preparation of control probes that are designed to fragment sequences that do not carry ESPs.

d) Fabrication of the micro-arrays. As described earlier by spotting either the ESP fragments or the oligonucleotide probes.

Genetic analysis of human individuals

For the micro-array two samples are prepared from each individual, a control sample and a CpG enzyme digested sample, using a cocktail of the 4 CpG

enzymes. Each sample will be prepared by combing the PCR amplification products of a series of amplification reactions using 50 ESP primer pairs each. The control sample utilizes Rox-labeled PCR primers, while Fau-labeled PCR products are used for the digested sample.

The two samples are then mixed and hybridized to the spotted micro-arrays. The hybridization signals of the control probes are used to normalize the signal intensities in both samples. For each of the ESP probes the ratio of the normalized Fam and Rox signal intensities is calculated. A ratio of 1 corresponds to, a ratio of 0.5 corresponds to heterozygous S+/S- and a ratio of <0.1 corresponds to homozygous S+.

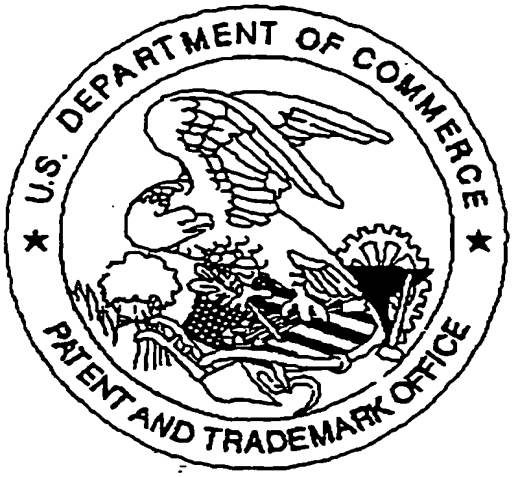
A similar assay is done pretreated DNA in which the control and the digested sample DNAs are both first digested with PacI and BfaI, ligated to the appropriate adapters and amplified with PacI and BfaI primers. This assay will give a better quantitative result.

The foregoing examples are illustrative of the invention and are not intended to be limiting. All of the references cited herein are incorporated by reference.

PRINT OF DRAWING AS ORIGINALLY FILED

United States Patent & Trademark Office

Office of Initial Patent Examination – Scanning Division



Application deficiencies were found during scanning:

☐ Page(s)_____ of No Declaration were not present for scanning.
(Document title)

☐ Page(s)_____ of _____ were not present for scanning.
(Document title)

☐ Scanned copy is best available.

Figure 1. Graphic representation of target fragments produced by consecutive cleavage with a rare cutter and a frequent cutter restriction enzyme. The type I fragments are shown as ~~~. The type II fragments are shown as solid lines

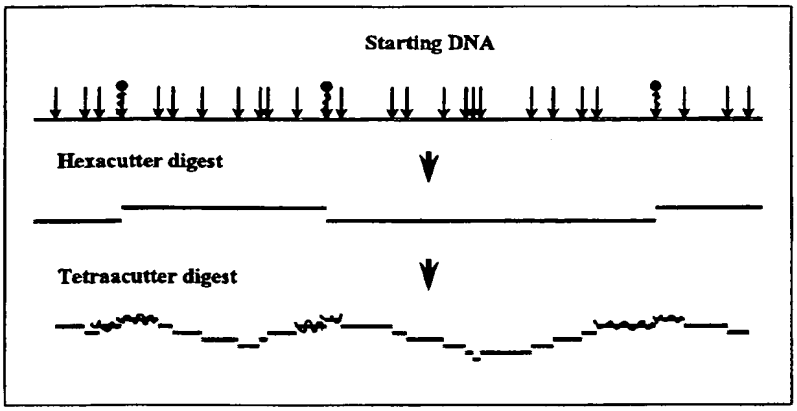


Figure 2. Graphic representation of the different types of mutations that can be detected. The upper panel depicts the target fragments produced by cleavage with a rare cutter and a frequent cutter, and the amplified fragments obtained thereof after treatment with a probing enzyme. The lower panel shows the different types of mutations affecting recognition sites.

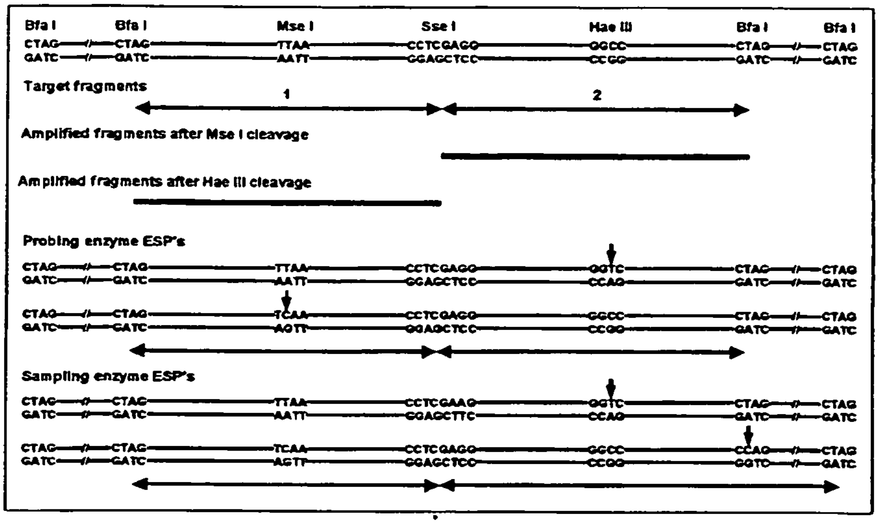


Figure 3

MseI selected ESP fragments **Ecotype Columbia**	**Tsp509I selected ESP fragments** **Ecotype Columbia**
MseI selected ESP fragments **Ecotype Landsberg**	**Tsp509I selected ESP fragments** **Ecotype Landsberg**

Figure 4 Hybridization patterns obtained on the four different micro-arrays: (1) ecotype Columbia digested with MseI, (2) ecotype Landsberg digested with MseI, (3) ecotype Columbia digested with Tsp509I and (4) ecotype Landsberg digested with Tsp509I. The green color results from the hybridization of the uncleaved fragment only, while the yellow color results from the simultaneous hybridization of the cleaved and the uncleaved fragment (combination of the green and the red fluorescent dyes).